XLQA: A Benchmark for Locale-Aware Multilingual Open-Domain Question Answering

Keon-Woo Roh[1], **Yeong-Joon Ju**[1], **Seong-Whan Lee**[1]
[1]Department of Artificial Intelligence, Korea University
{ro_keonwoo, yj_ju, sw.lee}@korea.ac.kr

Abstract

Large Language Models (LLMs) have shown significant progress in Open-Domain Question Answering (ODQA), yet most evaluations focus on English and assume locale-invariant answers across languages. This assumption neglects the cultural and regional variations that affect question understanding and answer, leading to biased evaluation in multilingual benchmarks. To address these limitations, we introduce XLQA, a novel benchmark explicitly designed for locale-sensitive multilingual ODQA. XLQA contains 3,000 English seed questions expanded to eight languages, with careful filtering for semantic consistency and human-verified annotations distinguishing locale-invariant and locale-sensitive cases. Our evaluation of five state-of-the-art multilingual LLMs reveals notable failures on locale-sensitive questions, exposing gaps between English and other languages due to a lack of locale-grounding knowledge. We provide a systematic framework and scalable methodology for assessing multilingual QA under diverse cultural contexts, offering a critical resource to advance the real-world applicability of multilingual ODQA systems. Our findings suggest that disparities in training data distribution contribute to differences in both linguistic competence and locale-awareness across models. https://github.com/ro-ko/XLQA

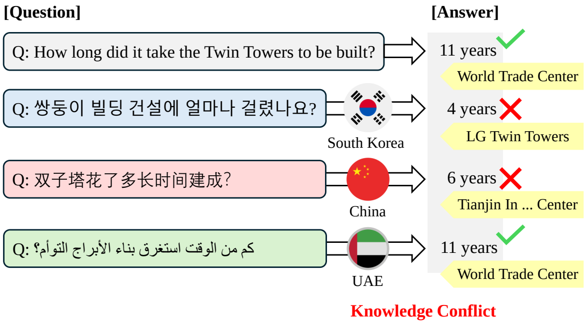


Figure 1: **Knowledge conflict in multilingual ODQA.** Although all versions of the question aim to ask how long it took to build the "Twin Towers", different languages elicit different answers based on locale-variant understanding. While English and Arabic refer to the World Trade Center (11 years), Korean and Chinese interpret "Twin Towers" as the LG Twin Towers and Tianjin IFC, respectively.

1 Introduction

Open-domain question answering (ODQA) aims to generate accurate and natural language answers to user queries without explicit domain constraints or provided context (Chen et al., 2017; Karpukhin et al., 2020). Recently, large language models (LLMs) (Brown et al., 2020; Anil et al., 2023; Workshop et al., 2022) have driven significant advances in ODQA by generating correct and natural answers. Despite strong advances in ODQA, most efforts have focused on English, leaving multilingual capabilities that remain relatively underexplored. This gap underscores the need for multilingual ODQA benchmarks that assess performance across languages (Maxutov et al., 2024).

To evaluate multilingual ODQA systems, existing benchmarks, such as MLQA (Lewis et al., 2020), MKQA (Longpre et al., 2021), and TyDiQA (Clark et al., 2020), are typically constructed by translating or aligning parallel questions across multiple languages. These benchmarks have the locale-agnostic assumption that both the meaning of a question and its correct answer remain constant across linguistic boundaries. However, this assumption overlooks variations in meaning that arise naturally from distinct cultural or regional contexts (Lin and et al., 2021; Liu et al., 2024; Zhang et al., 2023).

Recent benchmarks such as CaLMQA (Arora et al., 2025), NativQA (Hasan et al., 2025b), and BLEnD (Myung et al., 2024) attempt to overcome this limitation by constructing culturally grounded questions independently for each language. While these approaches provide valuable insights into

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28809–28821
November 4-9, 2025 ©2025 Association for Computational Linguistics

culture-specific reasoning, they do not directly ensure cross-lingual consistency, making systematic comparison across languages more challenging.

This issue introduces evaluation bias (Talat et al., 2022; Woo et al., 2023) by penalizing responses that are correct within specific regional or cultural contexts. For instance, as illustrated in Fig. 1, the answer to the question "How long did it take the Twin Towers to be built?" differs depending on which entity the question refers to: the World Trade Center in the U.S. or the LG Twin Towers in South Korea. Multilingual question requires the locale-variant references that arise from differing cultural contexts and background knowledge, not merely generating translated answers. In addition, relying on naive translation to construct multilingual benchmarks risks semantic drift, where subtle shifts in meaning occur due to inadequate contextual grounding (Yu et al., 2023). While human annotation can mitigate the drift, it is costly, labor-intensive, and difficult to scale across many languages and cultures (Pandey et al., 2022).

To address these challenges, we propose **XLQA**, a benchmark explicitly constructed to evaluate multilingual ODQA systems under locale-sensitive conditions. XLQA consists of 3,000 seed questions in English, each paired with a reference answer and language-specific supporting evidence. These questions are extended to eight languages (English, Korean, Arabic, Hebrew, Japanese, Russian, Vietnamese, and Simplified Chinese), resulting in 24,000 high-quality evaluation items. We design XLQA to assess whether multilingual ODQA systems can handle locale-sensitive variation by explicitly distinguishing between two types of questions: those whose correct answers remain consistent across languages (locale-invariant), and those whose answers vary depending on regional or linguistic context (locale-sensitive).

To construct this benchmark at scale, we apply a back-translation-based filtering method to identify and remove translations that exhibit potential semantic inconsistencies. Then, we generate locale-aware answers for each semantically consistent multilingual question by producing responses based on language-specific evidence curated for each locale with an LLM. These generated answers that semantically differ from the original English answer is categorized as a potentially locale-sensitive question. Human annotators examine each candidate instance to verify the answer's correctness and the relevance of the supporting evidence. This approach enables scalable multilingual QA dataset creation with limited human involvement, ensuring quality through selective verification rather than full manual annotation.

To demonstrate the effectiveness of this pipeline, we evaluate five multilingual LLMs on our benchmark, such as GPT-4.1 (Achiam et al., 2023), Qwen-3 (Zheng et al., 2025), Gemma-3 (Team et al., 2025), LLaMA-3.1 (Grattafiori et al., 2024), and EXAONE (Research et al., 2024) under standard evaluation metrics, including exact match and F1 score. Our analysis reveals that, despite strong zero-shot and multilingual capabilities, these models frequently fail to produce appropriate answers to locale-sensitive questions. We observe differences in both language proficiency and locale-specific knowledge across models, shaped by the distribution of language data used during training. These findings highlight the limitations of existing multilingual QA benchmarks and underscore the importance of explicitly modeling cultural context in evaluation. We summarize our contributions as follows:

- We introduce the first systematic framework for evaluating locale-aware correctness in multilingual QA, directly addressing the cultural insensitivity and English-centric assumptions embedded in prior benchmarks.
- We propose a scalable method for identifying and validating questions whose correct answers vary across regions, producing a benchmark of 3,000 high-quality question–answer–evidence triples annotated for locale sensitivity.
- We provide empirical evidence that current multilingual LLMs struggle with locale-grounded question answering, revealing a critical gap in their real-world applicability.

2 Related Work

2.1 Multilingual ODQA Benchmarks

In recent years, numerous multilingual question answering (QA) benchmarks have been proposed to evaluate the performance of multilingual language models. Prominent examples include MLQA (Lewis et al., 2020), XQuAD (Artetxe et al., 2020), TyDiQA (Clark et al., 2020), and MKQA (Longpre et al., 2021), which are widely used to compare model performance across different languages.

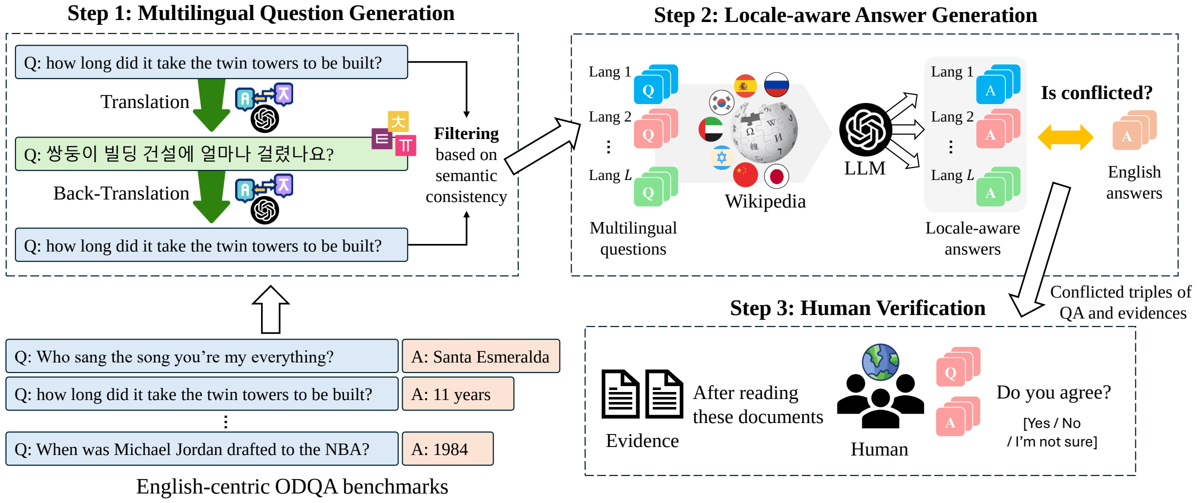

Step 1: Multilingual Question Generation
Q: how long did it take the twin towers to be built?
Translation
Q: 쌍둥이 빌딩 건설에 얼마나 걸렸나요?
Back-Translation
Q: how long did it take the twin towers to be built?
Filtering based on semantic consistency
Step 2: Locale-aware Answer Generation
Lang 1
Lang 2
Lang L
Multilingual questions
Wikipedia
LLM
Lang 1
Lang 2
Lang L
Locale-aware answers
Is conflicted?
English answers
Conflicted triples of QA and evidences
Step 3: Human Verification
Evidence
After reading these documents
Human
Do you agree?
[Yes / No / I'm not sure]
Q: Who sang the song you're my everything?
A: Santa Esmeralda
Q: how long did it take the twin towers to be built?
A: 11 years
Q: When was Michael Jordan drafted to the NBA?
A: 1984
English-centric ODQA benchmarks

Figure 2: **The overall pipeline for constructing the XLQA benchmark.** The process consists of three stages: (1) **Multilingual Question Generation** generates multilingual questions based on seed questions from existing QA datasets. (2) **Locale-Aware Answer Generation** uses LLM to generate locale-aware answers. (3) **Human Verification** verifies the answers with supporting evidence. The Output is a high-quality, locale-aware multilingual QA dataset.

MLQA and XQuAD are constructed by translating English question–answer pairs into multiple target languages, and rely on the assumption that the translated versions are semantically equivalent to the original. This approach enables direct comparison across languages but may overlook subtle linguistic or cultural differences that affect answer validity. In contrast, TyDiQA enhances linguistic diversity by collecting questions written natively in each language by fluent speakers, rather than relying on translation. However, it still assumes a single ground-truth answer per question within each language, potentially limiting its ability to capture within-language ambiguity or region-specific variation. MKQA takes a different approach by sourcing questions from anonymized Google Assistant logs, reflecting more natural, real-world user queries. These questions are then manually translated into 26 languages for open-domain question answering. While these benchmarks provide a foundation for measuring multilingual capabilities and cross-lingual consistency, they largely focus on surface-level correctness and lexical alignment. As such, they fall short of evaluating model performance in scenarios that require the understanding of cultural context or locale-specific knowledge.

2.2 Multilingual QA Evaluation Bias and Fairness

Recent works (Singh et al., 2024; Hasan et al., 2025a) have examined these issues from multiple perspectives. Singh et al. (2024) evaluates language models across culturally diverse multiple-choice questions. They show that performance varies substantially across languages and regions, indicating potential cultural bias. Hasan et al. (2025a) introduces a dataset of naturally occurring, culturally aligned queries in multiple languages. Their findings highlight the limitations of translation-based benchmarks in capturing region-specific information needs.

Bias is observed in model behavior across languages with differing resource levels, particularly in the form of stereotypical associations related to gender, profession, or ethnicity. Buscemi et al. (2025) proposes an automated evaluation framework to assess such social biases across both high- and low-resource languages. The study finds that these biases, such as associating certain professions more frequently with specific genders, tend to be more pronounced in low-resource settings, where training data is sparser and less balanced.

Similarly, Zulaika and Saralegi (2025) adapts the English-centric BBQ benchmark to Basque in order to investigate bias propagation in a typologically distant language. Their findings reveal that common bias mitigation strategies developed for English, such as data augmentation or counterfactual training, often fail to generalize effectively to underrepresented languages, underscoring the need for culturally and linguistically tailored approaches. These studies point to the need for evaluation methods that distinguish between culturally invariant and culturally dependent questions, and that reflect the diversity of real-world language use above high-resource settings.

2.3 Evaluation for LLM-as-judges

LLM-as-judge is a generative evaluator paradigm where LLMs are trained to produce an evaluation (natural language explanation and judgment) given the original user input, evaluation protocol (rules and criteria for evaluation), and model responses as input. JudgeLM (Zhu et al., 2025) formalizes this approach as a generative evaluation framework and demonstrates that LLM-based judges can approximate human evaluations in tasks such as reasoning and factual correctness. PandaLM (Wang et al., 2024) further investigates the reliability and robustness of LLM-based evaluators by comparing their preferences across model outputs with those of human annotators.

3 XLQA Dataset

To rigorously evaluate multilingual ODQA in locale-sensitive contexts, we introduce XLQA, a new benchmark constructed through our multi-stage pipeline. This pipeline consists of three steps: multilingual question generation, locale-aware answer generation, and human verification, as illustrated in Fig. 2.

3.1 Step 1: Multilingual Question Generation

We begin by collecting high-quality English seed questions from the test sets of existing ODQA benchmarks, such as MKQA (Longpre et al., 2021), MLQA (Lewis et al., 2020), and HotpotQA (Yang et al., 2018), to ensure alignment with our evaluation objectives. To refine the seed pool, we first remove duplicate entries based on an exact match of either the question or the answer. We then filter out unanswerable questions or those lacking a reference answer, as such items prevent meaningful comparison of locale-sensitive responses. This filtering process results in the exclusion of 28.4% of the initial seed questions.

For the refined seed questions, we generate multilingual questions translated into diverse target languages by utilizing GPT-4.1 as an Oracle Language Model (OracleLM), which refers to a theoretical upper-bound model that is assumed to know the correct answer, often used to estimate performance ceilings and analyze the gap between idealized and real-world behavior (Achiam et al., 2023; Chen et al., 2024). GPT-4.1 demonstrates strong performance in translation quality and contextual understanding, making it a suitable choice for ensuring the reliability of the generated multilingual questions. To ensure semantic consistency across the translated questions, we apply a back-translation filtering step. Each translated question is first back-translated into English. Then, the resulting back-translated version is compared against the original English question using the LLM-as-judge framework. GPT-4.1 is prompted to determine whether the two questions are semantically equivalent, providing a binary "yes/no" judgment. If any of the eight language translations are judged as inconsistent (i.e., the model outputs "no"), the entire question is discarded from the dataset. By discarding questions with inconsistent translations, this back-translation filtering step plays a crucial role in eliminating translation artifacts and mitigating cross-lingual meaning drift.

3.2 Step 2: Locale-Aware Answer Generation

To construct QA pairs that capture locale-specific variation, we generate candidate answers for the multilingual questions obtained in the previous step. For each input question, GPT-4.1 is prompted to generate an answer that reflects the locale associated with the language in which the question is written. For questions that are not sensitive to locale, the model is prompted to provide a general, culturally neutral answer. We leverage a retrieval-augmented generation (RAG) framework in which GPT-4.1 is connected to a web search component. This setup enables the model to generate answers grounded in verifiable external sources, providing both the response and its corresponding evidence. The retrieval process prioritizes authoritative sources, with a preference for Wikipedia. In case that relevant information is not found on Wikipedia, the system falls back to reputable news outlets.

As a post-processing step, we discard any QA pairs in which the generated reference lacks a valid URL or does not include reliable source indicators such as the keywords "wikipedia" or "news". This filtering ensures that all retained answers are grounded in verifiable and trustworthy sources. This approach offers an efficient alternative to human annotation by enabling scalable, high-quality data generation while maintaining contextual relevance and answer verifiability.

3.3 Step 3: Human Verification

All candidate triples flagged for answer conflict are subjected to human verification. Annotators are provided with the question, answer, and supporting evidence for each language. They are asked to determine whether the answer is correct and supported by the evidence. This process yields a high-quality set of QA-evidence triples, each labeled as either locale-invariant or locale-sensitive. To ensure consistency and reduce annotation noise, we adopt a majority voting scheme across three annotators per instance. Only instances where at least two annotators agree on both correctness and sensitivity labels are retained; otherwise, the item is discarded. Statistics on annotator agreement rates after voting are provided in Appendix Table 7.

4 Dataset Analysis

4.1 Dataset Statistics

Our benchmark consists of 3,000 question–answer–evidence triples across eight languages: English, Korean, Arabic, Hebrew, Japanese, Russian, Vietnamese, and Simplified Chinese. Each English-origin question is translated into the target languages and paired with answers and evidential support adapted to the cultural or linguistic context of the target locale.

On average, questions contain 17–40 tokens depending on the language, while answers remain short (4–6 tokens). A total of 24,000 QA instances were created, including 3,000 in English and 21,000 across the seven other languages.

4.2 Consistency Filtering Results

To ensure semantic consistency across translations, we applied a back-translation-based filtering pipeline. QA pairs with substantial semantic shifts, such as changes in named entities, factual scope, or temporal modifiers, were flagged and removed. In total, 10.8% of the generated multilingual instances were discarded through this process.

We observed that the majority of the filtered instances involved mistranslations of culturally specific terms or reinterpretations of ambiguous expressions that altered the intended meaning. These cases were particularly prevalent in Arabic and Hebrew, where semantic drift often resulted from incorrect rendering of proper nouns and idiomatic language. Table 5 summarizes the number of discarded instances per language following the consistency filtering process.

4.3 Conflict Detection

A conflict is defined as a case where at least one language provides an answer that is semantically inconsistent with the English reference, under the assumption that such variation is due to regional knowledge or interpretation. For each question, we collected answers across all languages and compared them using string normalization and embedding-based semantic similarity. Questions exhibiting divergence in meaning, rather than surface expression, were manually validated as locale-sensitive. Among the 3,000 source questions, 2,356 (73.9%) were categorized as locale-sensitive, based on the presence of conflicting answers in at least one language. Table 5 presents the distribution of conflicts across languages. Arabic and Hebrew displayed the highest proportion of conflicts, while Japanese and Vietnamese showed comparatively lower divergence.

5 Benchmark Evaluation

We conduct a series of experiments to evaluate multilingual LLM performance on our locale-aware QA dataset. Our goal is to assess how well current models handle both locale-invariant and locale-sensitive questions, and to quantify the limitations of existing evaluation protocols when applied to culturally or regionally diverse inputs.

5.1 Experimental Setup

We evaluate five widely used large language models with multilingual capabilities: GPT-4.1, Qwen 3, Gemma 3, LLaMA 3.1, and EXAONE. These models vary in architecture, size, and pretraining corpora, representing a broad range of capabilities in multilingual understanding and generation.

All models are evaluated in a zero-shot QA setting without fine-tuning. For each QA pair, the model generates an answer using a consistent prompting format adapted for the language. We apply two evaluation metrics:

- **Exact Match (EM)**: A binary metric that assigns 1 if the predicted answer exactly matches any of the reference answers, and 0 otherwise:

$$\text{EM} = \begin{cases} 1, & \text{if prediction} = \text{reference} \\ 0, & \text{otherwise} \end{cases}$$

- **F1 Score**: Measures the token-level overlap between the predicted and reference answers.

Lang	Oracle LM		Gemma3 12B		Qwen3 14B		LLaMA3.1 8B		EXAONE 7.8B	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
en	89.11	90.97	**43.26**	**52.68**	**40.73**	**49.43**	**40.38**	**50.56**	**31.44**	**39.64**
ar	87.86	90.05	18.54	23.62	11.83	19.30	8.53	16.92	3.98	6.04
he	88.30	90.46	20.05	24.83	11.04	16.08	11.86	16.60	5.52	7.20
ja	88.45	92.50	22.81	45.10	19.74	44.03	9.10	37.73	7.34	26.22
ru	87.83	89.54	28.52	35.20	17.67	27.97	14.53	24.18	7.41	9.91
ko	86.73	88.29	22.18	26.56	15.44	19.91	11.55	15.68	15.81	20.32
zh_cn	**89.68**	**93.41**	16.22	37.91	26.39	47.57	11.58	36.48	7.66	25.22
vi	89.39	91.19	36.55	44.77	26.83	39.34	26.45	38.38	10.95	14.70
Avg.	88.42	90.80	26.02	36.33	21.21	32.95	16.75	29.57	11.26	18.65

Table 1: Results of the base models on the XLQA benchmark using EM and F1 scores.

Lang	Gemma3 12B				Qwen3 14B				EXAONE 7.8B			
	Non-Conflict		Least-Conflict		Non-Conflict		Least-Conflict		Non-Conflict		Least-Conflict	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
en	59.09	72.25	37.69	45.79	64.02	75.36	32.51	40.28	53.67	65.28	23.60	30.59
ar	37.06	46.26	12.01	15.64	27.80	40.54	6.20	11.80	8.90	12.21	2.25	3.86
he	38.63	47.18	13.50	16.94	23.71	32.16	6.58	10.41	9.63	12.49	4.07	5.33
ja	41.16	**67.03**	16.34	37.36	41.03	65.30	12.22	36.53	15.28	38.24	4.54	**21.98**
ru	47.41	57.02	21.86	27.50	30.69	49.38	13.07	20.42	10.95	15.48	6.15	7.95
ko	39.35	46.06	16.13	19.69	31.05	38.11	9.93	13.49	**30.93**	**38.48**	**10.48**	13.91
zh_cn	27.32	56.12	12.31	31.49	**50.54**	**71.87**	17.87	**39.00**	15.16	37.15	5.01	21.01
vi	**51.62**	63.79	**31.24**	**38.07**	41.40	61.34	**21.69**	31.58	18.05	24.30	8.45	11.31
Average	42.70	56.96	20.13	29.06	38.78	54.26	15.01	25.44	20.32	30.45	8.07	14.49

Table 2: EM and F1 scores of Gemma3 12B, Qwen3 14B, and EXAONE 7.8B under different conflict levels.

It is computed as the harmonic mean of precision and recall: **F1 Score** measures the token-level overlap between the prediction and the reference answer. It is computed as the harmonic mean of precision and recall:

$$\text{Precision} = \frac{|\text{Prediction} \cap \text{Reference}|}{|\text{Prediction}|} \quad (1)$$

$$\text{Recall} = \frac{|\text{Prediction} \cap \text{Reference}|}{|\text{Reference}|} \quad (2)$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

We evaluate both locale-invariant and locale-aware settings.

5.2 Main Results

(1) Performance gap between English and other languages. Table 1 presents the performance of five LLMs on the XLQA benchmark. While English achieves the highest scores across all models, performance on other languages drops, particularly for those involving culturally diverse or underrepresented regions such as Arabic, Hebrew, Korean, and Vietnamese. This suggests that despite multilingual pretraining, current models struggle to generalize locale-aware reasoning beyond high-resource languages like English.

(2) Performance degradation on culturally sensitive questions. Table 2 offers a more granular view by separating questions into *non-conflict* and *least-conflict* subsets. Here, we define a question as exhibiting *least conflict* when at least one of the language-specific responses differs semantically from all other responses. This categorization captures cases where locale-sensitive variation arises across languages, allowing us to directly measure the challenge posed by culturally grounded knowledge. The results show a consistent and substantial performance drop across all models when faced with locale-sensitive questions. This highlights that answering such questions effectively requires not only understanding the language but also retaining culturally grounded knowledge specific to each region. Interestingly, models trained with a regional focus tend to perform better on conflict questions in their respective languages. For example, EXAONE achieves the highest conflict F1 score on Korean and QWEN3 on Chinese. While exact language-wise pretraining proportions are not publicly disclosed, these results suggest that higher exposure to specific locale-language data during pretraining enables models to better handle culturally nuanced inputs in that region.

5.3 Prompt Sensitivity

We examine the impact of prompt design using Qwen3 across four variants: **EN** (English prompt) and **EN-LOC** (English with locale emphasis).

Table 4 shows that prompts with explicit locale guidance (EN-LOC) improve accuracy, especially for culturally sensitive languages like Arabic and Korean. However, over-conditioning can sometimes lead to stereotype-driven outputs. While EN-LOC prompts generally improve performance, the degree of improvement varies significantly across languages. The gains are especially pronounced in Japanese (+25.03), Chinese (+17.42), and Korean (+7.58), suggesting that locale-specific grounding is particularly beneficial in languages with strong locale reference frames.

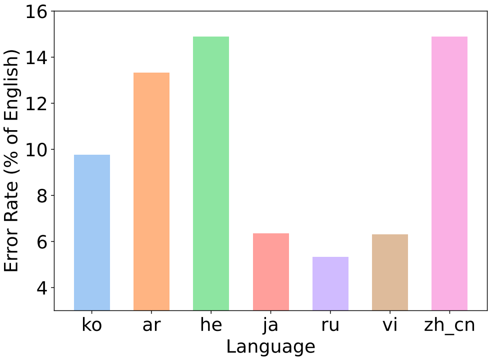


Figure 3: Comparison of translation error rates between naive translation and our back-translation pipeline.

5.4 Ensuring Semantic Consistency in Multilingual Questions

A back-translation-based filtering helps identify and remove mistranslations that may introduce unintended meaning shifts during naive machine translation. As shown in Figure 3, our back-translation pipeline significantly reduces translation error rates across most languages, particularly in Arabic, Hebrew, and Chinese languages that often exhibit greater semantic divergence from English. By improving the alignment between original and translated questions, this filtering step enhances the overall quality and reliability of locale-sensitive evaluation.

5.5 Categorization of Conflict-Inducing Questions

To better understand the sources of semantic divergence across languages, we manually categorize a subset of conflict-inducing questions based on the nature of the discrepancy observed in answers. This typology enables a more fine-grained analysis of the types of ambiguity and regional variability that arise in multilingual QA.

We categorize conflict-inducing questions into four types. These include *Entity Conflict*, *Factual Conflict*, *Cultural Reference*, and *Ambiguous Question*. **Entity Conflict** refers to cases where the referent entity varies across locales due to differing popularity or interpretation, such as entertainers or sports figures. **Factual Conflict** includes questions grounded in historical or statistical facts that may be represented differently depending on regional data sources. **Cultural Reference** covers instances involving awards, media, or events where local recognition or framing differs. Finally, **Ambiguous Question** includes vague or broadly interpretable queries that elicit culturally biased or interpretive responses.

Table 3 summarizes each conflict type along with representative subtopics, example questions, and the number of instances observed in our annotated subset. Entity-related conflicts were the most frequent, accounting for 1,032 questions, followed by Cultural References and Factual Conflicts. This distribution highlights the significant role of culturally grounded knowledge and localized salience in generating cross-lingual answer variability.

6 Conclusion

In this work, we identify a critical gap in existing multilingual QA benchmarks, the lack of consideration for locale-specific knowledge and culturally valid answer divergence. While prior evaluations assume semantic equivalence and a single correct

Conflict Type	Subtopics (Categories)	Representative Questions	Conflict Count
Entity Conflict	Music, TV actors, Sports players	*Who sang Oh What a Night?*, *Who played TJ on Head of the Class?*, *Who is the coach for the Toronto Raptors?*	1032
Factual Conflict	Geography, Political history, Team records	*How many states does the Rocky Mountains cover?*, *When was the last time the Lakers made the playoffs?*	431
Cultural Reference	TV show winners, Music awards, Famous media	*Who won America's Got Talent in 2015?*, *Who has the most Grammys?*	512
Ambiguous Question	Religion, Social media, General trivia	*Who wrote the Book of Lamentations?*, *Who has the most Instagram followers?*	381

Table 3: Conflict-inducing questions categorized by conflict type, with subtopics and representative examples.

Lang	EN	EN-LOC
en	48.03	49.43
ko	12.33	**19.91**
ar	11.93	19.30
he	16.37	16.08
ja	19.00	**44.03**
ru	16.41	**27.97**
vi	33.37	**39.34**
zh_cn	30.15	**47.57**
Overall	23.45	**32.95**

Table 4: Performance (F1 score) across languages under different prompting strategies on Qwen3.

Lang	Conflicted Answers	Conflict Rate (%)
ar	1471	46.2%
he	1413	44.3%
ja	1044	32.8%
ru	963	30.2%
ko	1188	37.3%
zh_cn	1242	39.0%
vi	909	28.5%
At Least One Conflict	2356	73.9%

Table 5: Language-wise distribution of answer conflicts in the XLQA benchmark.

	en	ar	he	ja	ko	ru	zh_cn	vi
Avg Question Length	37	33	31	26	22	40	17	38
Avg Answer Length	5	5	5	8	4	5	6	5

Table 6: Average question and answer lengths across languages (rounded to nearest integer).

answer across languages, our analysis shows that this assumption fails in questions involving cultural or regional context. To address this, we propose a method for constructing locale-aware evaluation subsets that allow for valid answer variation across languages. Our approach combines translation consistency checks and prompt-based answer divergence detection to identify culturally sensitive questions. We demonstrate that such questions are not rare, and that standard evaluation protocols may underestimate the capabilities of multilingual models in diverse linguistic settings. This work calls for a shift in multilingual QA evaluation toward frameworks that are not only linguistically fair but also culturally grounded.

Limitations

Our evaluation may be inherently bounded by the capabilities of the proprietary large language models (LLMs) accessed via API. Since these models serve as oracle systems for translation and answer generation, their performance imposes an upper bound on the quality and diversity of our data. To mitigate potential issues arising from translation artifacts or inconsistencies, we applied a semantic consistency filtering step using back-translation and LLM-as-judge comparison to ensure that the generated multilingual questions preserve the meaning of the original seed questions. Additionally, due to computational resource constraints, we were unable to include larger-scale open-source multilingual models that require substantial local infrastructure. To compensate for this limitation, we evaluated a diverse set of models—both proprietary and open-source—covering a range of capabilities and linguistic domains, and conducted all evaluations under a unified framework to ensure comparability. Future work could expand this line of research by integrating scalable open-source multilingual models in controlled environments and broadening the linguistic and regional scope of the evaluation.

7 Acknowledgment

This work was partly supported by the Institute of Information & Communications Technology

Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (Artificial Intelligence Graduate School Program (Korea University) (No. RS-2019-II190079), No. IITP-2025-RS-2024-00436857 (Information Technology Research Center (ITRC))) and Artificial Intelligence Star Fellowship Support Program to Nurture the Best Talents (IITP-2025-RS-2025-02304828)).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: A family of highly capable multimodal models. corr, abs/2312.11805, 2023. doi: 10.48550. *arXiv preprint ARXIV.2312.11805*.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11772–11817.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4623–4637.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901.

Alessio Buscemi, Cédric Lothritz, Sergio Morales, Marcos Gomez-Vazquez, Robert Clarisó, Jordi Cabot, and German Castignani. 2025. Mind the language gap: Automated and augmented evaluation of bias in llms for high-and low-resource languages. *arXiv preprint arXiv:2504.18560*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1870–1879.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics (TACL)*, 8:454–470.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025a. Nativqa: Multilingual culturally-aligned natural queries for llms.

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025b. NativQA: Multilingual culturally-aligned natural query for LLMs. In *Findings of the Association for Computational Linguistics (ACL)*, pages 14886–14909.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7315–7330.

Xi Victoria Lin and et al. 2021. Calmqa: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1829–1841.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 2016–2039.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transac-*

tions of the Association for Computational Linguistics (TACL), 9:1389–1406.

Akylbek Maxutov, Ayan Myrzakhmet, and Pavel Braslavski. 2024. Do LLMs speak Kazakh? a pilot evaluation of seven models. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK)*, pages 81–91.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 78104–78146.

Rahul Pandey, Hemant Purohit, Carlos Castillo, and Valerie L. Shalin. 2022. Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. *International Journal of Human-Computer Studies*, 160:102772.

LG Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Stanley Jungkyu Choi, Yemuk Choi, Seokhee Hong, Yeonjung Hong, Junwon Hwang, Hyojin Jeon, and 1 others. 2024. Exaone 3.0 7.8 b instruction tuned language model. *arXiv preprint arXiv:2408.03541*.

Amanpreet Singh, Yujia Wang, Yulia Tsvetkov, and Percy Liang. 2024. Global-mmlu: Evaluating cultural and linguistic biases in multilingual language understanding. *arXiv preprint arXiv:2412.03304*.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The International Conference on Learning Representations (ICLR)*.

Tae-Jin Woo, Woo-Jeoung Nam, Yeong-Joon Ju, and Seong-Whan Lee. 2023. Compensatory debiasing for gender imbalances in language models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:55734–55784.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don‘t trust chatgpt when your question is not in english: A study of multilingual abilities and types of LLMs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7915–7927.

Xingyu Zheng, Yuye Li, Haoran Chu, Yue Feng, Xudong Ma, Jie Luo, Jinyang Guo, Haotong Qin, Michele Magno, and Xianglong Liu. 2025. An empirical study of qwen3 quantization. *arXiv preprint arXiv:2505.02214*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. JudgeLM: Fine-tuned large language models are scalable judges. In *The International Conference on Learning Representations (ICLR)*.

Muitze Zulaika and Xabier Saralegi. 2025. BasqBBQ: A QA benchmark for assessing social biases in LLMs for Basque, a low-resource language. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 4753–4767.

A XLQA Construction Details

A.1 Prompt Templates

We provide the full prompt templates used throughout the XLQA benchmark construction and evaluation pipeline. These include:

Translation prompts, used to generate multilingual versions of questions from English.

> Given the question: {question}, please translate it into {loc}. Just output the translated question only, with no comments or formatting.

Back-translation prompts, used to back-translate to English.

> Given the question: {translated_response_output_text}, please translate it back into English. Just output the translated question only, with no comments or formatting.

Consistency filtering prompts, used to verify semantic consistency across languages.

> Given the question: {question}, please check if the back translation: {back_translation_output_text} is correct. If it is correct, output "yes". If it is not correct, output "no".

Locale-aware answer generation prompts, which condition the model to generate region-specific answers if appropriate.

> Answer the following question based on the cultural context of a region where the {lang} language is primarily spoken. If the correct answer would vary depending on regional or cultural differences, return the version that best fits that local context. However, if the question concerns universal or culturally-neutral knowledge, provide the common or globally accepted answer instead. Respond with only the final answer in a single word or phrase. Do not explain or add anything else. Additionally, provide a brief evidence or source (e.g., a Wikipedia URL, news site, or cultural explanation) that supports the answer. The question is: {q}

Answer generation prompts for evaluation, which elicit general answers (EN) or, when relevant, region-specific ones (EN-LOC).

> **(EN) General Prompt**
>
> Answer the following question. Respond with only the final answer in a single word or phrase. Do not explain or add anything else.

> **(EN-LOC) Locale-aware Prompt**
>
> Answer the following question based on the cultural context of a region where the {lang} language is primarily spoken. If the correct answer would vary depending on regional or cultural differences, return the version that best fits that local context. However, if the question concerns universal or culturally-neutral knowledge, provide the common or globally accepted answer instead. Respond with only the final answer in a single word or phrase. Do not explain or add anything else.

A.2 Human Verification Agreements Ratio

A.3 Locale Sensitivity Annotation Guidelines

We define a question as *locale-sensitive* if its correct answer may differ depending on regional, cultural, or national context, even when the semantic intent of the question remains the same.

Annotators were instructed to mark a question as locale-sensitive if:

Regionally salient knowledge affects the expected answer (e.g., “most famous tower”).

Political, institutional, or cultural prominence varies by country or language group.

The question involves subjective norms or identity references (e.g., “national dish”, “popular leader”).

Borderline cases were resolved by majority voting across annotators with multilingual and regional backgrounds.

B Experimental Details

B.1 Models

We use the following models in our experiments:

- **Gemma3 12B**: Uses Gemma3 with 12B parameters. Licensed under **Apache 2.0 license.**.

- **Qwen3 14B**: Uses Qwen3 with 14B parameters. Licensed under the **Apache 2.0 license.**.

- **LLaMA-3.1 8B**: Has 8B parameters and is released under the **LLaMA 3 Community License Agreement**.

- **GPT-4.1**: These models are not open-source and are accessible only via API requests. They are governed by **proprietary licenses**.

- **Exaone 7.8B**: Uses Exaone with 7.8B parameters. Licensed under **EXAONE AI Model License Agreement**.

All the models set the temperature to 0.

B.2 Budget

We use the RTX A6000 GPU X 1 with 20 hours.

Language	Correctness (≥2/3)	Correctness (≥2/3)	Sensitivity (≥2/3)	Sensitivity (≥2/3)
English (en)	91.2%	98.5%	88.3%	96.7%
Korean (ko)	89.7%	97.4%	85.2%	95.9%
Arabic (ar)	86.4%	96.1%	80.5%	93.8%
Hebrew (he)	88.1%	97.0%	82.7%	94.6%
Japanese (ja)	90.5%	98.1%	87.0%	96.2%
Russian (ru)	87.9%	96.8%	84.1%	94.3%
Vietnamese (vi)	89.3%	97.9%	86.5%	95.7%
Chinese (zh_cn)	88.7%	97.5%	83.6%	94.8%
Average	**88.9%**	**97.4%**	**84.7%**	**95.3%**

Table 7: Annotator agreement rates by language. The table shows the percentage of instances where all three annotators (3/3) or at least two annotators (2/3) agreed on correctness and locale-sensitivity labels.

C Human Annotation

To verify the correctness and locale sensitivity of the model-generated answers, we conducted human annotation using Amazon Mechanical Turk (MTurk). For each language, we recruited **three independent annotators** who are native or proficient speakers of the respective target language to evaluate each QA-evidence triple. Annotators were presented with the original question, the model-generated answer, and its associated supporting evidence (e.g., URL or passage), and were instructed to assess as in Figure 4.

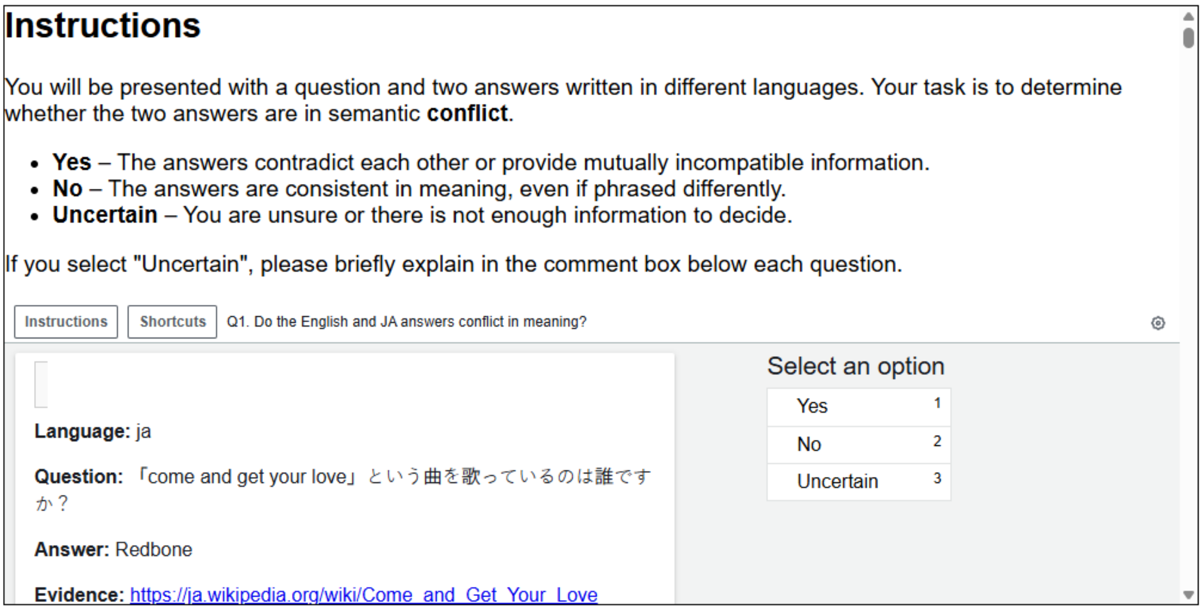
Each annotation instance was reviewed by three annotators. Final labels were determined via **majority voting**. Annotator agreement rates are summarized in Table 7.

All annotators were compensated at a rate of $5 per 100 questions, in line with MTurk compensation standards, and informed that their responses would be used for research purposes. No personally identifiable information was collected during the process. Tasks involving potentially sensitive content were manually reviewed and filtered prior to annotation to avoid harm or discomfort.

D Ethical Considerations

While XLQA promotes cultural inclusion in QA evaluation, locale-aware generation introduces ethical challenges. Prompts conditioned on locale risk overgeneralization or reinforcement of cultural stereotypes. We manually reviewed outputs for offensiveness and excluded instances containing bias or politically sensitive content.

Furthermore, hallucination in low-resource languages may amplify misinformation if locale grounding is weak. We recommend that future work incorporate human validation when deploying such systems in high-stakes settings.



Instructions

You will be presented with a question and two answers written in different languages. Your task is to determine whether the two answers are in semantic **conflict**.

- **Yes** – The answers contradict each other or provide mutually incompatible information.
- **No** – The answers are consistent in meaning, even if phrased differently.
- **Uncertain** – You are unsure or there is not enough information to decide.

If you select "Uncertain", please briefly explain in the comment box below each question.

Q1. Do the English and JA answers conflict in meaning?

Language: ja

Question: 「come and get your love」という曲を歌っているのは誰ですか？

Answer: Redbone

Evidence: https://ja.wikipedia.org/wiki/Come_and_Get_Your_Love

Figure 4: **Survey screenshot.** Interface shown to MTurk annotators during the human verification stage.